Forensic Network Traffic Analysis

Noora Al Khater
Department of Informatics
King's College London
London, United Kingdom
noora.al_khater@kcl.ac.uk

Richard E Overill
Department of Informatics
King's College London
London, United Kingdom
richard.overill@kcl.ac.uk

ABSTRACT

The nature of information in a network is volatile and dynamic, some precious evidence might be missed. The real-world situations need a quick classification decision before the flow finishes, especially for security and network forensic purposes. Therefore, monitoring network traffic requires a real-time and continuous analysis, to collect valuable evidence such as instant evidences that might be missed with post-mortem analysis (dead forensics). Network traffic classification is considered the first line of defence where a malicious activity can be filtered, identified and detected. In addition, it is the core component in evidence collection and analysis that uses filtered evidence and helps to reduce redundancy. However, most of the existing approaches that deal with collecting evidence from networks are based on post- mortem analysis. Therefore, this research investigates different classification techniques using Machine Learning (ML) algorithms, seeking to identify ways to improve classification methods from a forensic investigator standpoint.

KEYWORDS

Digital forensics; cyber security; network traffic classification; digital evidence; network analysis

1 INTRODUCTION

Analysing network traffic is considered a pro-active investigation in network forensics. Monitoring and classifying network traffic correctly can help organisations avoid a lot of harm, and provide them with an intelligible vision of traffic patterns to put a good incident plan response in place. In fact, organisations are looking for more effective methods to solve these problems that caused by increased in number of cyber-crimes that cause genuine harm.

There are great historical developments of classification techniques; however, they are not free from drawbacks, which can be outlined in two points. First, traditional techniques (port-based classifications) are unreliable because most of the applications use random or non-standard ports. Second, even payload-based classification is considered more accurate but this accuracy diminishes when dealing with encrypted traffic; furthermore, this technique has a highly expensive computational cost and may be thwarted by privacy policies because it inspects packet contents. These problems have driven the research community to shift into statistical and behavioral classification using Machine Learning (ML) techniques, which depend on analysis of application patterns that do not require payload inspection. Additionally, the increasing amount of traffic and transmission rates stimulate researchers to look for lightweight algorithms. And the persistence of application developers in inventing new ways to avoid filtering and detection mechanisms of traffic is another motivating factor.

This paper structured as follows: Section 2 defines digital forensics and provides background on the forensic model. Section 3 discusses the use of ML algorithms in the field of network traffic analysis and reviews the existing studies. Section 4 outlines the

challenges in the field of network traffic classification using ML. Section 5 illustrates the proposed method for forensic network traffic analysis. Section 6 summarises our research.

2 WHAT IS DIGITAL FORENSICS?

The Digital Forensics Research Workshop (DFRWS) [1] has defined digital forensics as: "The use of scientifically derived and proven methods toward the preservation, collection, validation, identification, analysis, interpretation, documentation, and presentation of digital evidence derived from digital sources for the purpose of facilitation or furthering the reconstruction of events found to be criminal, or helping to anticipate unauthorized actions shown to be disruptive to planned operations."

Network forensics is subdomain of digital forensics. It deals with the analysis of network traffic as proactive investigation, and this is different to other areas of digital forensics because it involves volatile information that requires real-time continuous analysis. Network forensics has been defined by DFRWS [1] as: "The use of scientifically proven techniques to collect, fuse, identify, examine, correlate, analyze, and document digital evidence from multiple, actively processing and transmitting digital sources for the purpose of uncovering facts related to the planned intent, or measured success of unauthorized activities meant to disrupt, corrupt, and or compromise system components as well as providing information to assist in response to or recovery from these activities."

2.1 Background on Forensic Model

For investigation in digital forensic science there are several models, which can be used. For a clear explanation we take a formal comprehensive approach from the DFRWS. The DFRWS model consists of sequential steps in the process of digital forensic analysis [1]. The steps help the practitioners and researchers to conceive of the situation to understand the direction of what they need to focus on. These steps can be seen in **Table 1.**

The linear approach begins with the process of identifying potential evidence. This involves the location of the digital evidence and how to determine the location. For example, in network forensics Intrusion Detection Systems (IDS) can be sources of digital evidence; IDS involves recognising and detecting unusual patterns of flow in the network traffic. The second step is preservation, which is a critical phase for increasing the possibility of a successful investigation. This process starts from acquiring, seizing, and preserving the evidence to create a digital image of the evidence and maintain the chain of custody. The following steps are collection, examination and analysis of the digital evidence to culminate in the final single presentation. Network forensics requires real-time collection of the digital evidence to avoid any missing critical information. However, most of the methods in the collection process are based on post- mortem analysis. The huge amount of collected data needs automated techniques to reduce redundancy, and consequently reduce the analysis time of the evidence [2]. Manual analysis of a complex network attack involving a large amount of data is impossible in a timely fashion, so an automated method is a basic element to identify, collect and link evidence with the criminal action. In addition, the massive amount of traffic needs a huge storage capacity, which costs a lot; this amount of data needs to be filtered to extract related data. Actually, the capture and analysis of network traffic is normally based on sniffing tools, which face difficulty and whose effectiveness diminishes when they deal with encrypted packets. However, analysis techniques based on pattern recognition using ML algorithms have proven promising results. Finally, the last stage of digital forensic investigation is presentation, which entails dealing with legal aspects of the

case and presenting the investigation's findings in the court.

Table 1. Digital Forensic Investigation Process [1]

Identification		**Preservation**	**Collection**	**Examination**	**Analysis**	**Presentation**
Event/Crime Detection	Case Management	Preservation	Preservation	Preservation	Preservation	Documentation
Resolve Signature	Imaging Technologies	Approved Methods	Traceability	Traceability	Expert Testimony	
Profile Detection	Chain of Custody	Approved Software	Validation Techniques	Statistical	Clarification	
Anomalous Detection	Time Synch.	Approved Hardware	Filtering Techniques	Protocols	Mission Impact Statement	
Complaints		Legal Authority	Pattern Matching	Data Mining	Recommended Countermeasure	
System Monitoring		Lossless Compression	Hidden Data Discovery	Timeline	Statistical Interpretation	
Audit Analysis		Sampling	Hidden Data Extraction	Link		
		Data Reduction		Spatial		

3 NETWORK TRAFFIC CLASSIFICATION USING ML TECHNIQUES

Despite of the fastest and simplest of the previous classification technique (port-based) in monitoring and reporting activity of network traffic, the unreliability of port-based classification has been proved in several published works. For instance, one study [3] stated that the port base analysis method can not classify 30- 70% of the internet traffic used in its work. Also, another investigation of the accuracy of port-based classification [4] demonstrated that the accuracy of this traditional technique, using the official IANA list, is not better than 70% of bytes. However, this classification is still used in situations where accuracy is not a critical matter.

The problems of port-based classification methods and the drawbacks of payload-based classification techniques have motivated researchers to find alternative ways to analyse network traffic using statistical and behavioural properties of the flow without relying on inspection of packet contents. ML techniques have shown promising results in analysing network traffic based on the extracted features of the flow.

3.1 Machine Learning Concepts

ML is a core branch of artificial intelligence. It is the science that makes machines able to obtain new information, develop new skills, assess and reorganize existing knowledge, and identify new information. In the early 1990s Shi [5] noted that the distinctive characteristic of intelligence is that machines have the ability to learn form experience automatically. Additionally, in 2000 Witten and Frank spotted that "Things learn when they change their behavior in a way that makes them perform better in the future" [6].

ML techniques are used in several applications such as medical diagnosis, search engines, marketing diagnosis, etc. ML algorithms were developed and their applications disseminated over various fields. The first remarkable work of ML techniques in the field of telecommunications networking was in 1990; this work was an attempt to get the most out of call completion in a circuit switcher [7]. 1994 saw the start of using ML algorithms to classify internet flow for intrusion detection [8]. This was the spark for a lot of research on applying ML techniques in network traffic classification.

3.2 Types of Machine Learning

There are four different types of Machine learning according to Witten and Frank [6]: supervised learning (classification), unsupervised learning (clustering), association, and numeric prediction. Supervised learning uses a number of pre-classified examples, to build classifier rules to identify unknown flow. Unsupervised learning automatically classifies network flow into groups (clusters) of instances that have the same properties without any kind of pre-guidance. Association learning explores the links among features. And in numeric prediction, the prediction result is the numeric volume, not a discrete class.

Most of the studies in the field of network traffic classification used supervised learning; a few studies, however, have applied unsupervised learning (clustering) or investigated the use of hybrid techniques (semi-supervised) in their analysis of network traffic.

3.2.1 Supervised Classification

Supervised classification techniques use pre-classified (pre-labelled) samples during the training phase to build a classifier (model) with a set of rules in order to classify new samples [9]. The information learnt can be presented in the form of classification rules, decision tree, flowchart, etc. This information can then be used to identify similar examples. Two main processes are inherent in supervised classification—namely, the training process and then the final step, the testing process. The testing phase is the following process where the model (classifier) is used to identify new flow. Supervised techniques are suitable for identifying specific types of applications. The effectiveness of these techniques is subject to a training phase (training set), because these methods of classification focus on forming the relationships of input/output.

There are a number of different ML algorithms that can be used in classification. Each algorithm is distinct and different in the construction process of the model and the output. Examples of these algorithms are the Nearest Neighbours (NN), Linear Discriminate Analysis (LDA) and Quadratic Discriminant Analysis (QDA) algorithms. The authors in [10], applied a statistical signature-based approach using these ML algorithms to classify IP traffic. Another study [11] used the supervised ML naive Bayes technique to classify network traffic. In the training phase they used 248 full flow-based features. They were able to achieve 65% flow accuracy in their classification results with a simple Naive Bayes technique. And they enhanced the results of the classification by using the Naive Bayes Kernel Estimation (NBKE) and Fast Correlation-Based Filter (FCBF) techniques, enhancing flow accuracy results to 95% overall. The results of this study were improved upon in [12] by applying the Bayesian neural network technique. The authors were able to achieve 99% accuracy when the classifier was trained and tested at the same time and 95% when the classifier was tested eight months later.

In [13], the authors applied Naive Bayes ML algorithm to classify network traffic based on features extracted from sub-flows instead of full flows. With this technique they were able to avoid the classifier to look for the start of the flow which could be missed. The authors in

[14] extended their previous study [13] by using Naive Bayes and decision-tree algorithms. They affirmed once again that performance of classifiers is enhanced when they are trained using sub-flows with the same datasets as in [13]. They also compared their results with the poor results of classification based on statistical features extracted from bi-directional flows.

Another work [15] used Genetic Algorithm (GA) for feature selection, applied three different algorithms (Naive Bayesian classifier with Kernel Estimation (NBKE), Decision Tree J48, and the Reduced Error Pruning Tree (REPTree)) and compared their classification results. The classification outcomes show that the performance of tree J48 and REPTree are better than NBKE with high accuracy in the results. This study mentioned the influence of using data from several sites for training and testing processes. Although the accuracy result is based on overall results, it is noteworthy that high accuracy is provided by the first 10 packets used in the classification. This study raises the question: If different applications were used, how different would the results be?

3.2.2 Unsupervised Classification

In contrast to supervised classification, unsupervised classification techniques automatically classify flow without any guidance, trying to find internalised heuristics from unlabelled data based on discovering patterns [16]. They do this by clustering instances that have relatively similar characteristics or are identical in terms of input data together in groups. Some of the instances are moved into only one group, which can be called a limited group. Other situations involve overlapping, when the same instance can be found in different groups.

The basic unsupervised techniques are the k-means algorithm, probability- based clustering, and incremental clustering [6]. Unsupervised technique was studied in [17] by applying the Expectation Maximisation (EM) algorithm [18] to classify network traffic based on features extracted from full flows. Using this technique, the researchers grouped applications into several groups based on characteristics. Their technique can be taken as an initial step in identifying unknown network traffic.

Other researchers, such as [19], proposed a method to classify TCP flow by applying unsupervised ML (Simple K-Means algorithm). Their proposed classification technique exploits features taken from the first packets of the TCP flow. This contrasts with the previous study [20], which used the features of full flows. This proposed approach allows a quick identification of the applications that flow over the network by examining the first few packets only. This classification method is based on the hypothesis that assumes the classifier can always know the beginning of every flow. This is not the case in reality, as in real-world network traffic the start of the flow might be missed. Therefore, the classifier needs to consider different conditions and scenarios; otherwise the capability of the classifier decrease in situations different than studied conditions.

3.2.3 Semi-Supervised Classification

Semi-supervised classification combines supervised and unsupervised techniques, taking advantage of both techniques by having the ability to detect new applications like unsupervised techniques, which reduces the performance degradation of supervised classification when dealing with situations different than studied conditions where new application emerges.

One of the significant pieces of research on semi-supervised techniques was conducted by [21]. The proposed classification method in this study started with supplying a clustering algorithm with a training dataset that combined labelled flows and unlabelled flows. The

provided labelled flows help a clustering algorithm to map clusters with labelled flows. The unknown flows are allocated to the closest cluster based on the distance metric they used. The initial results with a semi-supervised approach using the K-Means clustering algorithm is promising. Further details of the results can be found in [22].

4 RESEARCH CHALLENGES

After considering the most recent developments in field of network traffic analysis using ML, it is clear that there are some problems associated with the methodology that have been used to analyse and classify network traffic in most of the published works. However, there is still space for significant contributions. Real-world situations require prompt analysis-based decision-making before the network flow ends—especially in critical situations involving the identification and detection of security incidents, for example.

It is important to mention that the majority of the proposed techniques deteriorate when dealing with different conditions than ones for which they are best suited, like coping with a huge amount of network traffic (i.e. real situations). This means a lot of applications might not carry out classification and identification correctly; hence, valuable forensic evidence might be missed. Also, it is notable that most studies applied ML techniques using features extracted from full flows, whereas the results of classifying using sub-flows outperformed the results of studies using full flows. A very few studies, including [13, 23], investigated this. Indeed, for security analysis purposes the speed in identifying network traffic is also required as the accuracy of the results. The problem of classifying using full flows is that if the beginning of the flow is missed, performance of the classifier diminishes. In addition, losing the beginning of the packet is a very common scenario in real-world network traffic.

In point of fact and from a network forensics and security standpoint, delay, fragmentation, packet loss, and emergence of a new application are very important aspects that need to be considered in the performance evaluation of the classifier for more effective and realistic classification results, however most of the studies in this field have not investigated these issues. Furthermore, unsupervised techniques have not been used widely in classifying network traffic, despite the fact that these techniques have the ability to identify unknown applications if used properly or combined with supervised algorithms to enhance their effectiveness in analysing network traffic.

5 The PROPOSED TECHNIQUE

Taking account of these gaps in the field and the importance of correctly and effectively classifying network traffic as a kind of pro-active investigation, this research will investigate different network traffic classification techniques relying on the analysis of statistical flows using ML algorithms, without inspecting packet contents or using traditional techniques (port-based classification).

The sequential steps of our work as illustrated in **Figure 1.** The first step involves capturing network traffic to create an appropriate dataset for our investigation. The second step consists of extracting features, such as the duration and length of the packet and inter-packet arrival times, from the sub-flow. We chose to work with features extracted from sub-flow because we believe this is the fastest way to analysis network traffic instead of using full-flow especially for security and forensic aspects. The extracted features will be used to train the classifier in the supervised techniques stage, with the endpoint of analysing network traffic smoothly. There is the high expectation that the classifier would not be able to distinguish a new application properly, if we were to use several datasets other than that on which the classifier

was trained; this study looks to investigate this hypothesis by adding unseen examples to various datasets. The performance of the supervised classifier will be evaluated, based on its ability to identify unseen applications; this is an important aspect that needs to be considered, for more realistic scenario.

In the next stage, the objective will be to enhance the performance of the clarifier by combining supervised and unsupervised learning techniques. The aim in this stage will be to leverage the ability of unsupervised algorithms to identify unclassified applications. We will then evaluate the results of using both types of technique, to determine any differences in performance.

In the third stage, we will investigate the capability of unsupervised techniques to work independently. Through this work, we will look to find a way of improving the performance of classification techniques, without relying solely on the use of supervised techniques. Additionally, we look also to emphasize the importance of apply unsupervised techniques in network traffic analysis and investigate the ability of these techniques to distinguish new and unseen applications that could be malicious in nature. The idea behind this research is that we can evaluate the extent of their effectiveness in coping with real-world situations—situations that require both the prompt recognition of application type in the network before the flow ends, and the making of quick decisions. Such timeliness is critical, as many criminal incidents could be detected and prevented before they can inflict extensive damage. This impetus explains our motivation in choosing to apply unsupervised techniques to identify unseen applications in forensic network traffic analysis.

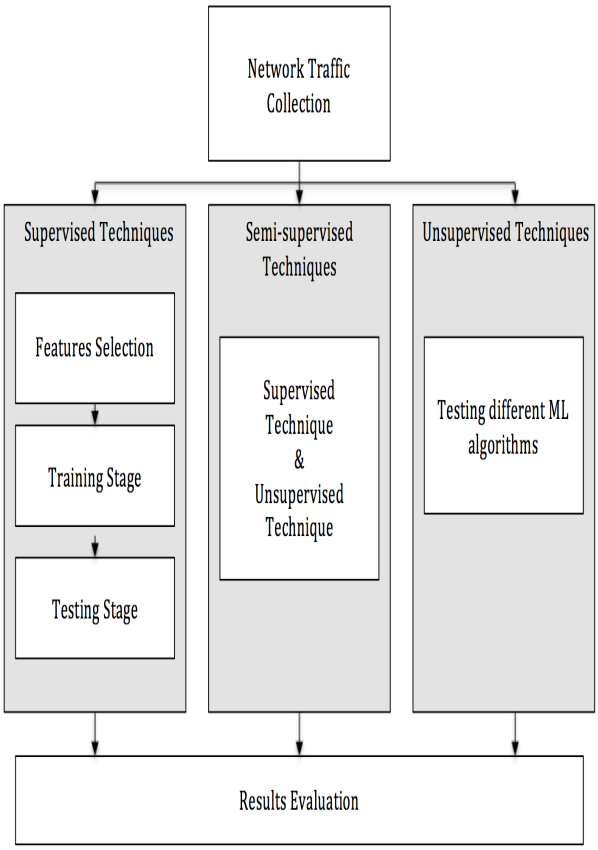

Figure 1. The Workflow of Forensic Network Traffic Analysis

6 CONCLUSION

Network traffic is a source of evidence that needs to be correctly classified immediately. Detection and response must occur prior to the attack. As a result, a real-time analysis of the application and automated classification is required to analyse the network, which is a valuable source of forensic evidence. Consequently, this provides structured forensic investigations of network security incidents. However, most of the evidence collection techniques for network forensics are rely on post- mortem analysis. Hence, a lot of precious evidence for investigation might be lost, which can lead to an inaccurate conclusion due to the weak evidence. In fact, to build a clear and strong case lawyers need more corroborating evidence, which calls for real-time collection.

Therefore, this research aims to identify methods that can improve classification techniques by using ML algorithms, and create a balance among accuracy, efficiency, and cost, because a large amount of network traffic is still unclassified.

REFERENCES

[1] Digital Forensics Research Workshop. (2001, November). A road map for digital forensics research 2001. Retrieved from http://www.dfrws.org

[2] W. Wang and T. Daniels. "Building Evidence Graphs for Network Forensics Analysis," in Proceedings of the 21st Annual Computer Security Applications Conference (ACSAC 2005). September 2005.

[3] A. Madhukar and C. Williamson, "A longitudinal study of P2P traffic classification," in 14th IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems, September 2006.

[4] A. W. Moore and K. Papagiannaki, "Toward the accurate identification of network applications," in Passive and Active Measurement, Boston, MA, March 2005, pp. 41–54.

[5] Z. Shi, Principles of Machine Learning. International Academic Publishers, 1992.

[6] I. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, 2nd ed., Morgan Kaufmann Publishers, 2005.

[7] B. Silver, "Netman: A learning network traffic controller," in Proceedings of the Third International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems, Association for Computing Machinery, 1990.

[8] J. Frank, "Machine learning and intrusion detection: Current and future directions," in Proceedings of the National 17th Computer Security Conference, Washington, D.C., October 1994.

[9] Y. Reich and J. S. Fenves, "The formation and use of abstract concepts in design," in Concept Formation: Knowledge and Experience in Unsupervised Learning, D. H. Fisher and M. J. Pazzani, Eds. Morgan Kaufmann, 1991.

[10] M. Roughan, S. Sen, O. Spatscheck, and N. Duffield, "Class-of-service mapping for QoS: A statistical signature-based approach to IP traffic classification," in Proceedings of ACM/SIGCOMM Internet Measurement Conference (IMC) 2004, Taormina, Sicily, Italy, October 2004.

[11] A. Moore and D. Zuev, "Internet traffic classification using Bayesian analysis techniques," in ACM International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS) 2005, Banff, Alberta, Canada, June 2005.

[12] T. Auld, A. W. Moore, and S. F. Gull, "Bayesian neural networks for Internet traffic classification," IEEE Transactions on Neural Networks, no. 1, pp. 223–239, January 2007.

[13] T. Nguyen and G. Armitage, "Training on multiple sub-flows to optimise the use of machine-learning classifiers in real-world IP networks," in Proceedings of the IEEE 31st Conference on Local Computer Networks, Tampa, FL, pp. 369–376, November 2006.

[14] T. Nguyen and G. Armitage, "Synthetic sub-flow pairs for timely and stable IP traffic identification," in Proceedings of the Australian Telecommunication Networks and Application Conference, Melbourne, Australia, December 2006.

[15] J. Park, H.-R. Tyan, and K. C.-C. J. Kuo, "GA-based Internet traffic classification technique for QoS provisioning," in Proceedings of the 2006 International Conference on Intelligent Information Hiding and Multimedia Signal Processing, Pasadena, CA, December 2006.

[16] H. D. Fisher, J. M. Pazzani, and P. Langley, Concept Formation: Knowledge and Experience in Unsupervised Learning. Morgan Kaufmann, 1991.

[17] A. Mcgregor, M. Hall, P. Lorier, and J. Brunskill, "Flow clustering using machine learning techniques," in Proceedings of the Passive and Active Measurement Workshop 2004, Antibes Juan-les-Pins, France, pp. 205–214, April 2004.

[18] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," J. Roy. Stat. Soc., vol. 30, no. 1, pp. 1–38, 1997.

[19] L. Bernaille, R. Teixeira, I. Akodkenou, A. Soule, and K. Salamatian, "Traffic classification on the fly," ACM SIGCOMM Comput. Commun. Rev., vol. 36, no. 2, pp. 23–26, 2006.

[20] S. Zander, T. Nguyen, and G. Armitage, "Automated traffic classification and application identification using machine learning," in IEEE 30th Conference Local Computer Networks 2005, Sydney, Australia, pp. 250–257, November 2005.

[21] J. Erman, A. Mahanti, M. Arlitt, I. Cohen, and C. Williamson, "Semisupervised network traffic classification," ACM International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS) Perf. Eval. Rev., vol. 35, no. 1, pp. 369–370, 2007.

[22] J. Erman, A. Mahanti, M. Arlitt, I. Cohen, and C. Williamson. "Offline/realtime traffic classification using semi-supervised learning." Perf. Eval., vol., 64 nos. 9–12, pp. 1194–1213, 2007.

[23] T. Nguyen, G. Armitage, P. Branch, and S. Zander, "Timely and continuous machine-learning-based classification for interactive IP traffic," IEEE/ACM Trans. Netw., vol. 20, no. 6, pp. 1880–1894, December 2012.